Simulations Showing Normal Distribution for Sample Mean and Sample Variance of Exponential Distribution

Phillip Rowe, 9/12/2019

Simulation Exercise

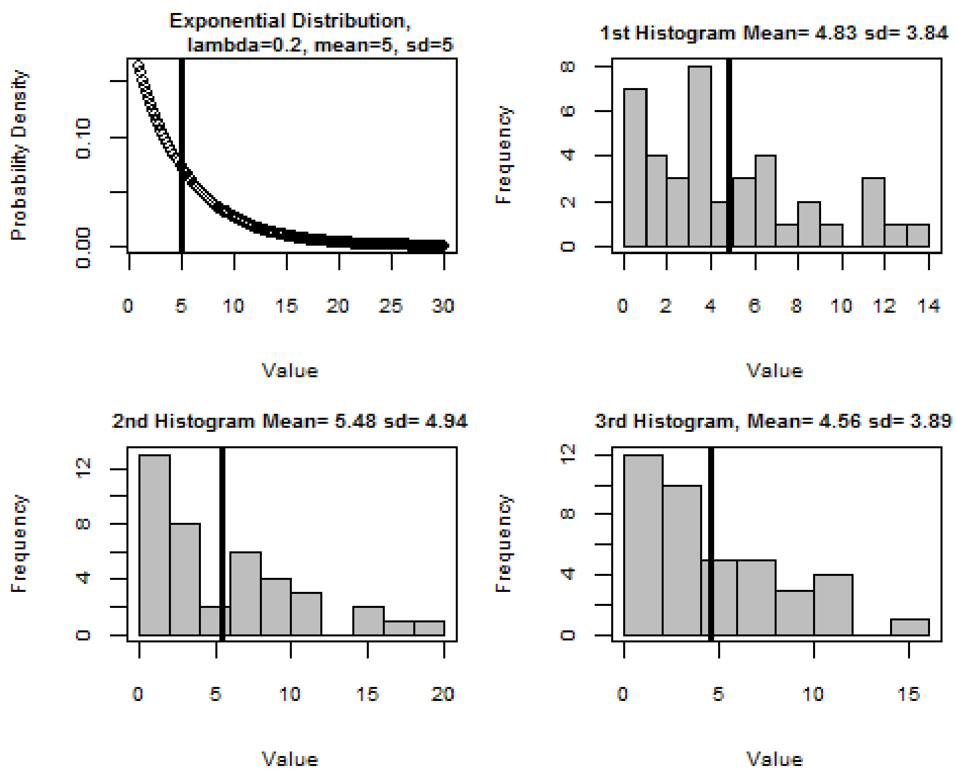
Overview

We will demonstrate several points of the Central Limit Theorem: 1) that the sample mean and sample variance estimate the theoretical mean and variance of the exponential probability distribution and 2) that the sample mean and sample variance have normal probability distributions with expected value equal to the expected value of the sampled exponential distribution, 3) the variance of the sample mean random variable is approximated by the exponential distribution's variance divided by the number of samples used to create each sample mean (or equivalently, the standard deviation of the sample mean distribution is approximately the exponential's standard deviation divided by the square root of the number of samples used to create each sample mean, i.e., 40 samples).
Simulations

The exponential distribution is given by f(x)=lambda*exp(-lambda*x), for x>=0, and its expected value (mean) and standard deviation are both equal to 1/lambda (see Figure 1 for plot of PDF). Samples can be generated in R using `rexp(n, lambda)`. We set lambda to 0.2 (mean = standard deviation = 5; variance = sd^2 = 25). In the code below, we create a dataframe of 1000 rows x 40 columns where each row is 40-samples of an exponential distribution. We then plot in Figure 1 three histograms of the first three rows of the dataframe.

```
library("dplyr");  library("plyr"); library(lubridate); library('reshape2')

lambda=0.2
set.seed(10)
# store 1000 simulations of 40 samples of distribution in a dataframe
alldata = NULL
for (i in 1 : 1000) alldata = rbind(alldata, rexp(40,lambda))
#calculate 1000 means, one for each of 40-sample rows
allmns<-apply(alldata,1,mean); bigmean<-round(mean(allmns),2)
     #calculate 1000 standard deviations, one for each of 40-sample rows
allsds<-apply(alldata,1,sd); bigsd<-round(mean(allsds),2)

par(mfrow=c(2,2),mai = c(.7, 0.7, 0.3, 0.1))
x<-seq(1,30,.2)
y<-lambda*exp(-(lambda*seq(1,30,.2)))
# a plot of the Exponential distribution with expected value 1/lambda
plot(x, y, ylab="Probability Density", xlab="Value",  main="Exponential Distribution,
          lambda=0.2, mean=5, sd=5",cex.main=.75, cex.lab=.75, cex.axis=.75)
abline(v=5,lwd=3)

mean1<-round(mean(alldata[1,]),2); sd1<-round(sd(alldata[1,]),2)
title<-paste("1st Histogram Mean=",mean1,"sd=",sd1)
hist(alldata[1,],10,ylab="Frequency",xlab="Value", main=title,cex.main=.75,
     col="gray",cex.lab=.75, cex.axis=.75)
```

```
abline(v=mean1,lwd=3); box()

mean2<-round(mean(alldata[2,]),2); sd2<-round(sd(alldata[2,]),2)
title<-paste("2nd Histogram Mean=",mean2,"sd=",sd2)
hist(alldata[2,],10,ylab="Frequency",xlab="Value", main=title,cex.main=.75,
     col="gray",cex.lab=.75, cex.axis=.75)
abline(v=mean2,lwd=3); box()

mean3<-round(mean(alldata[3,]),2)
sd3<-round(sd(alldata[3,]),2)
title<-paste("3rd Histogram, Mean=",mean3,"sd=",sd3)
hist(alldata[3,],10,ylab="Frequency",xlab="Value", main=title,cex.main=.75,
     col="gray",cex.lab=.75, cex.axis=.75)
abline(v=mean3,lwd=3); box()
```

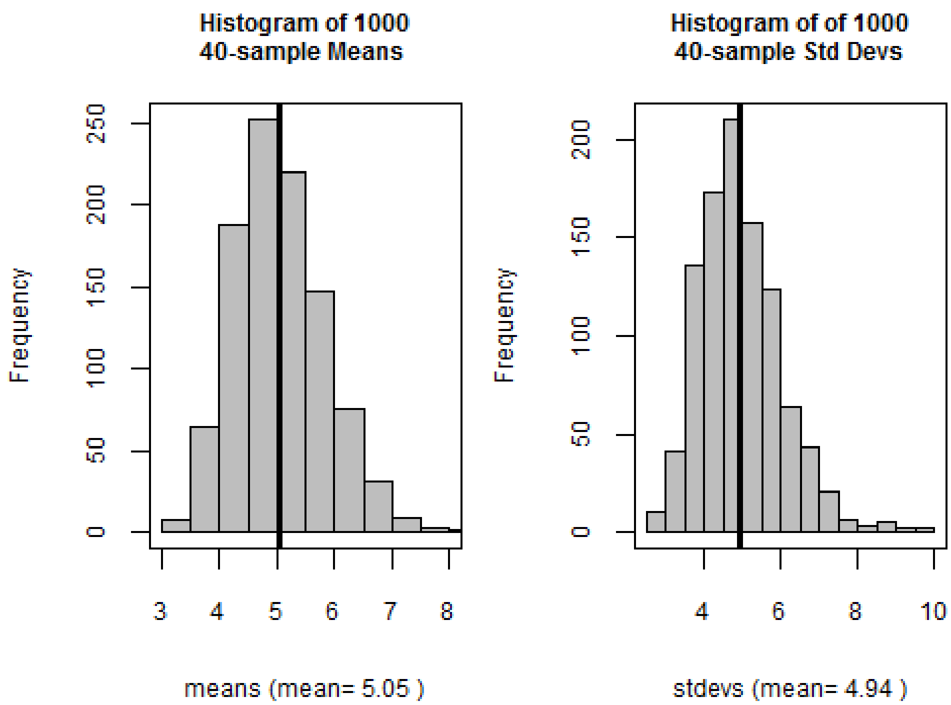


Sample Mean vs. Theoretical Mean; Sample Variance versus Theoretical Variance

Due to the random nature of the sampling and the relatively small sample size (n=40), some of these histograms approximate the PDF in the upper left corner better than others, and thus the sample mean and sample standard deviation shown on each histogram are only somewhat close to the theoretical values. That said, over the 1000 simulations, we calculated all the 40-sample means and standard deviations (stored in `allmeans` and `allsds`), and plotted their histograms, which appeared to be normal distributions in Figure 2. The mean of all 1,000 sample means was 5.05, and the mean of all 1,000 sample standard deviations was 4.94, both much closer to the theoretical values of the exponential distribution where lambda=0.2. Note that instead of comparing the variances we compared the standard deviations, which is just the square root of variance.

The final lines of code show the third and final point of this report's Overview. The variance (or standard deviation) of the 1000 sample means is very close to its theoretical value: the exponential's variance divded by the number of samples. We repeat this calculation using the standard deviation.

```r
# now compare histograms of means and variances, which should look Gaussian
par(mfrow=c(1,2),mai = c(1, .8, .7, 0.1))
xlabel<-paste("means (mean=",bigmean,")")
hist(allmns, xlim=c(3,8), col="gray", ylab="Frequency",xlab=xlabel,
     main="Histogram of 1000\n40-sample Means ",cex.main=.75, cex.lab=.75, cex.axis=.75)
abline(v=bigmean,lwd=3); box()

hist(allsds,col="gray", xlab=paste("stdevs (mean=",bigsd,")"),ylab="Frequency",
     main="Histogram of of 1000\n40-sample Std Devs", cex.main=.75, cex.lab=.75,
     cex.axis=.75)
abline(v=bigsd,lwd=3); box()
```




```r
var(allmns)

## [1] 0.6372544

# theoretical value
25/40

## [1] 0.625

# doing the same with the standard deviation
# the standard deviation of the 1000 sample means, which are random variables
sd(allmns)

## [1] 0.7982821

# should be close to its theoretical value of the exponential's standard dev / sqrt of
# number of samples used to generate each mean (i.e., 40)
5/sqrt(40)

## [1] 0.7905694
```